Machine Translation

Dr. Rajendran, S
Amrita Vishwa Vidyapeetham
Coimbatore
rajushush@gmail.com

Machine translation

- **Machine translation is the study of designing systems that translate from one human language into another.**
- This is a hard problem, since processing natural language requires work at several levels, and **complexities and ambiguities arise** at each of those levels.
- Some pragmatic approaches can be used to tackle these issues, leading to extremely useful systems.

Issues in Machine Translation

- Machine translation (and natural language processing in general), is a difficult problem. There are **two main reasons**, which are related.
- The first reason is that natural language is highly ambiguous**.** The ambiguity occurs at all levels – *lexical,* syntactic, semantic and pragmatic.
- A given word or sentence can have more than one meaning.
- For example, the word '*party*' could mean a political entity, or a social event, and deciding the suitable one in a particular case is crucial to getting the right analysis, and therefore the right translation.

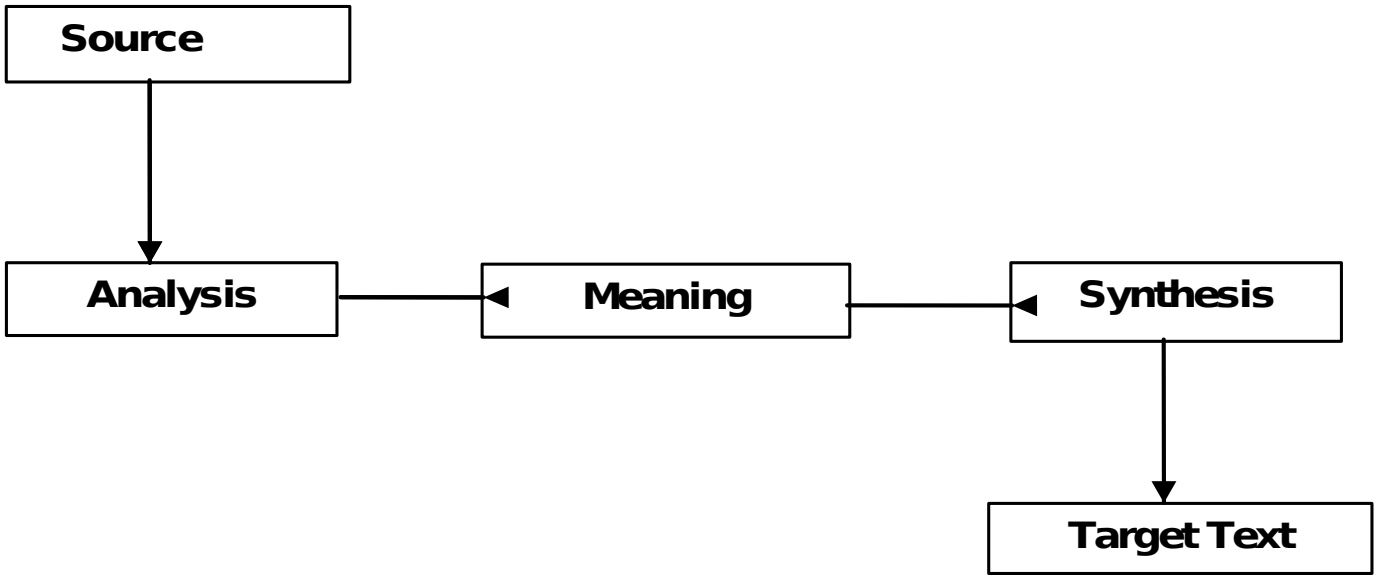
Issues in Machine Translation

- The second reason is that when humans use natural language, they use an enormous amount of common sense, and knowledge about the world, which helps to resolve the ambiguity.
- For example, in *"He went to the bank, but it was closed for lunch",* we can infer that *'bank'* refers to a financial institution, and not a river bank, because we know from our knowledge of the world that only the former type of bank can be closed for lunch.
- To get MT systems to exhibit the same kind of world knowledge in an unrestricted context requires a lot of effort.

The translation process

The translation process may be stated as:

1. Decoding the meaning of the source text; and
2. Re-encoding this meaning in the target language.



Different Categories of Machine Translation Systems

The three categories of machine translation systems are:

- Machine Aided Human Translation
- Human Aided Machine Translation
- Fully Automated Machine Translation

Various Approaches to Machine Translation

- Linguistic or Rule Based Approaches
  - Direct Approach
  - Interlingua Approach
  - Transfer Approach
- Non-Linguistic Approaches
  - Dictionary Based Approach
  - Corpus Based Approach
    - Example Based Approach
    - Statistical Approach
- Hybrid Approach

Difference between Rule based and SMT

- Rule based MT
  - requires bilingual dictionary
  - requires morphological analyzer for source language
  - requires morphological generator for target language
  - requires source side parser for handling the syntax; Target parser is not required
- Statistical MT
  - requires huge parallel corpora
  - recent advancements in SMT allows incorporation of linguistic information such as Part of Speech (PoS) and Morphology.

Linguistic or Rule Based Approaches

- Rule based approaches requires a lot of linguistic knowledge during the translation and
- so it uses grammar rules and computer programs which will be helpful in analysing the text for determining grammatical information and features for each and every word in the source language,
- translating it by replacing each word by the target equivalent or word that have the same context in the target language.

Linguistic or Rule Based Approaches

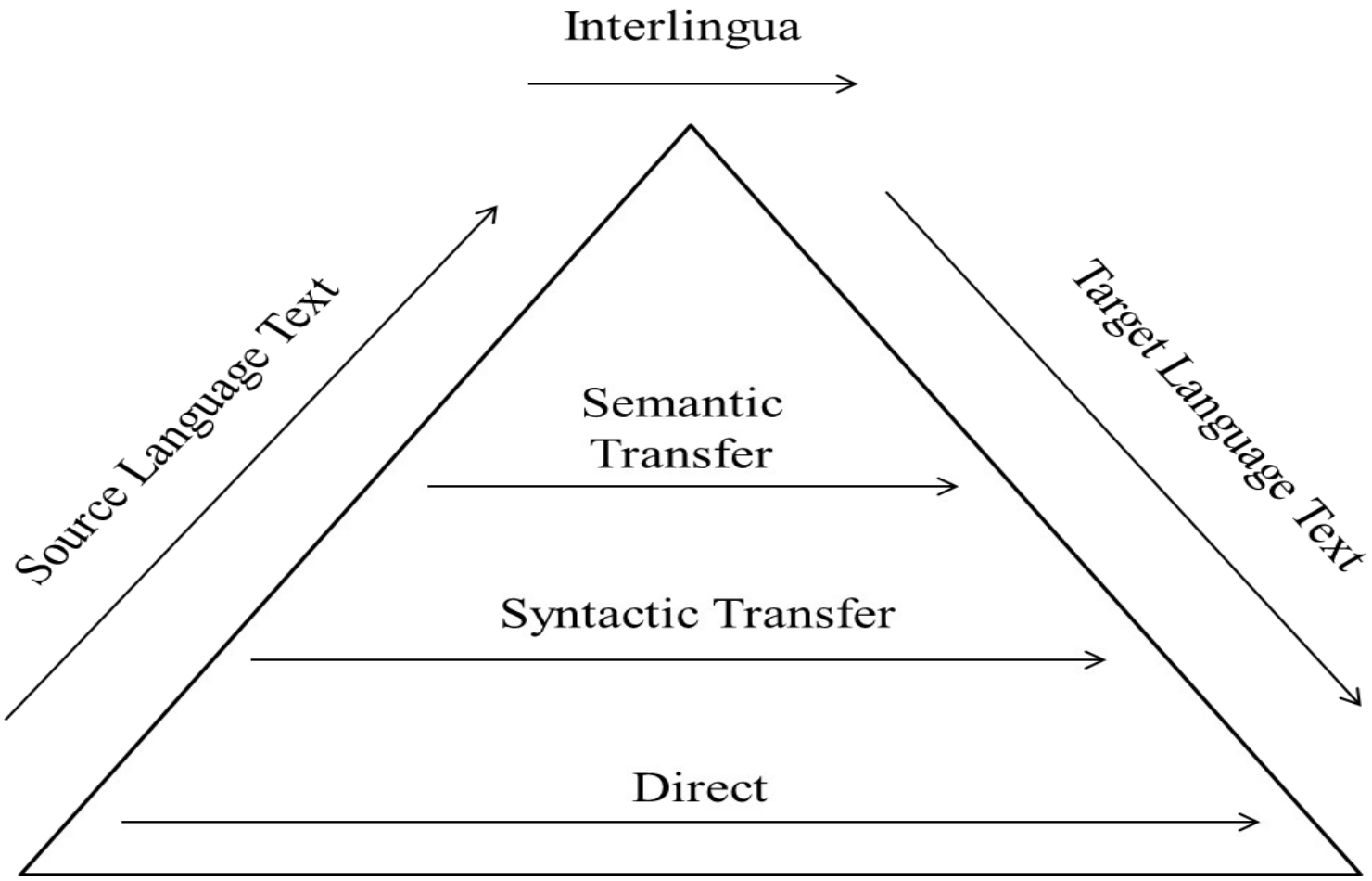
- Rule based approach is the principal methodology that was developed in machine translation.
- Linguistic knowledge is required in order to write the rules for this type of approaches. These rules will play a vital role during the different levels of translation.
- The benefit of rule based machine translation method is that it can intensely examine the sentence at its syntax and semantic levels.
- There are complications in this method such as the prerequisite of vast linguistic knowledge and the vast number of rules needed to maintain the balance between source and target languages

Linguistic or Rule Based Approaches

The three different approaches that require linguistic knowledge are as follows:

- Direct Approach
- Interlingua Approach
- Transfer Approach

The Vauquios Triangle of Machine Translation



Components of a typical MT System

- We can divide the machine translation task into two or three main phases –
  - The system has to first **analyse** the source language input to create some internal representation.
  - It then typically manipulates this internal representation to **transfer** it to a form suitable for the target language.
  - Finally, it **generates** the output in the target language.
- Thus a typical MT system contains components for analysis, transfer and generation.

Levels Of Analysis Of Source Language

- Lexical level (or word level)
- Syntactic level (or sentence level)
- Semantic level (meaning level)
- Discourse level
- Pragmatic level

Components of a typical MT System

- **ANALYSIS COMPONENT – ANALYSER**
  - Tokanizer : Tokenizes the sentence into words/ tokens
  - □□ □□ □□□□□ □□□□□ □□□□. ->
    □□
    □□
    □□□□□
    □□□□□
    □□□□
    .

Components of a typical MT System

- Morphological analyzer: It analyses words into morphemes and gives a meaning representation

  □□□□□□- □□□□□□□□,

  □□□□□- □□□□□□□,

  □ - □□□

- POS tagger: Assign Parts-of-speech to the tokens / words

  □□/PN □□/DET □□□□□/ADJ □□□□□/N □□□□/V

- Syntactic parser: Analyses sentences into phrases or chunks.

  [□□/PN]NP [□□/DET □□□□□/ADJ □□□□□/N]NP [□□□□/V]VP

Components of a typical MT System

- TRANFER COMPONENT – TRANSFER SYSTEM
  - It transfers parsed source language syntactic structure into target language syntactic structure

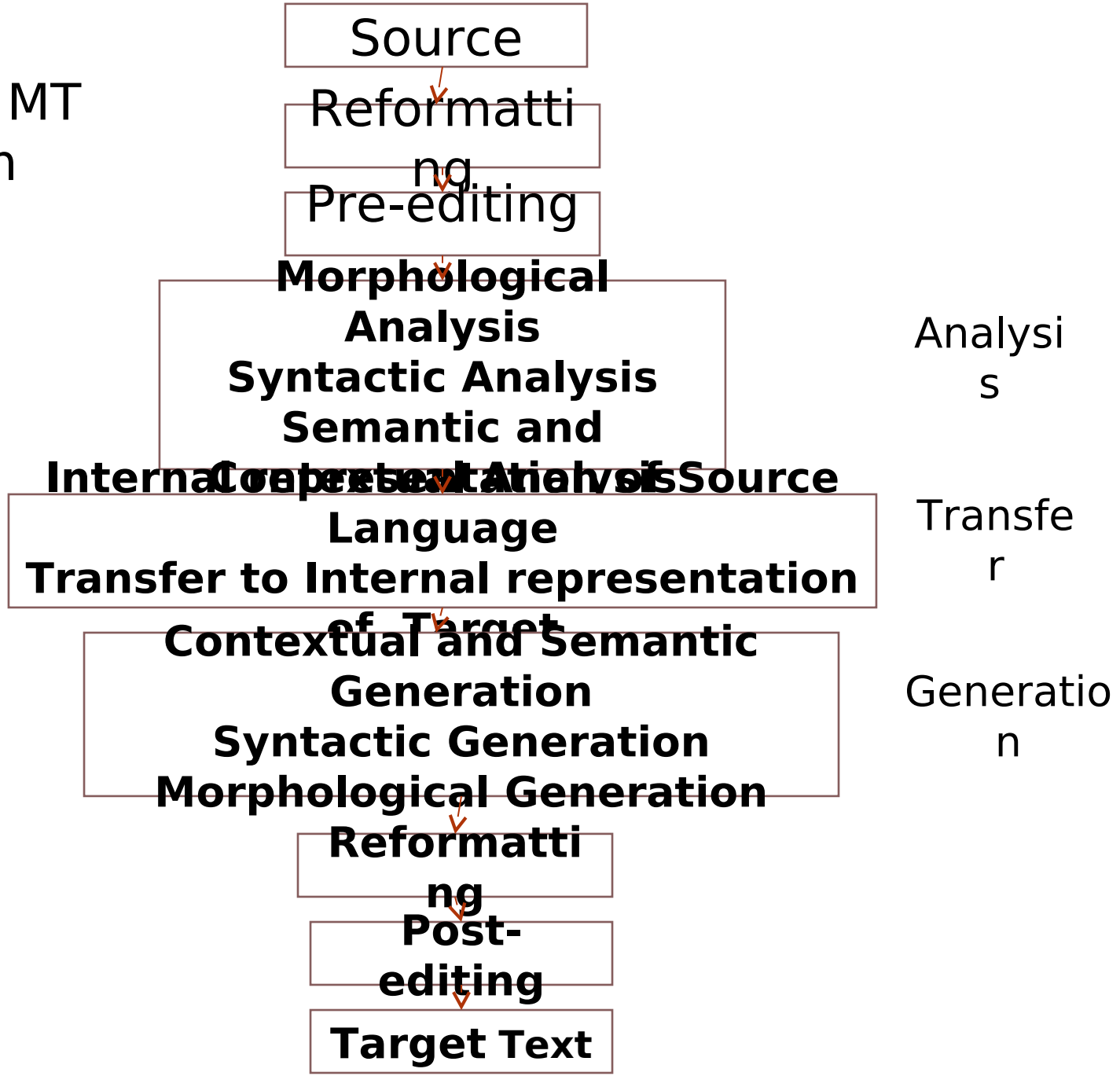
    he read a book -> he a book read ->
  - वह एक किताब पढ़ी
- GENERTION - GENERATOR
  - Generates target language words form source language morphological information.

    पुस्तक + PLURAL -> पुस्तकें

    आ + PAST -> आया
  - Generates target language sentence from target language's morphological and syntactic information.

    he read a book -> वह एक किताब पढ़ी .

A typical MT system



Morphological processing

- Morphology is the study of how individual words are built up from smaller meaning bearing units called morphemes.
- There are two types of morphological processing: Morphological analysis and Morphological generation.
- The system should understand that '*foxes*' is a plural form of '*fox*', '*mice*' is a plural form of '*mouse*' and '*butterflies*' is a plural form of '*butterfly*'
- Morphological processing is crucial for Indian languages which are morphologically rich.
- Morphotactics, the arrangement of morphemes in a word need to be taken into considering while processing words.
- Cascade of Finite State Transducers(FST) have been successfully used to analyze or generate word forms.
- Software tools: PC-KIMMO, AT&T FSM Library, OpenFST from Google

Shallow Parsing

- Abney (1991) is credited with being the first to argue for the relevance of shallow parsing, both from the point of view of psycholinguistic evidence and from the point of view of practical applications.
- His own approach used hand-crafted cascaded Finite State Transducers to get at a shallow parse.

Shallow Parsing

Typical modules within a shallow parser architecture include the following:

- Part-of-Speech Tagging: Given a word and its context, decide what the correct morphosyntactic class of that word is (noun, verb, etc.). POS tagging is a well-understood problem in NLP, to which machine learning approaches are routinely applied.
- Chunking: Given the words and their morphosyntactic class, decide which words can be grouped as chunks (noun phrases, verb phrases, complete clauses, etc.)
- Relation Finding: Given the chunks in a sentence, decide which relations they have with the main verb (subject, object, location, etc.).

Part of Speech(POS) Tagging

- Parts-of-Speech (POS) tagging – the task of assigning appropriate POS tag to every word in a given text – is an important process used as a building block for several Natural Language Processing (NLP) tasks.
- A POS tagset usually defines the list of morphosyntactic categories that are applicable at the word-level to a specific language, though in some cases they may also include pure morphological or syntactic information.
- The early efforts in POS tagset design in 1970s, that resulted in tagsets such as UPenn, Brown and C5, mainly focused on tagsets for English which were mostly simple lists of tags corresponding to the morphosyntactic features, and varied greatly in terms of their granularity (Hardie, 2004).

Part of Speech(POS) Tagging

- Part of Speech (POS) tagging is the process of assigning a Part of Speech to words.
- Each word in a corpus is marked for lexical class/POS.

  Input text: “I missed the train”

  Output text: I/Pronoun missed/VBD the/DET train/Noun
- POS tagging solves ambiguities at word level.

  Ex. ‘book’ can be Verb or Noun, ‘object’ can be Verb or Noun.
- Rule based techniques as well as Stochastic techniques such as Hidden Markov Models (HMM), Maximum Entropy models, Support Vector Machines (SVM) and Conditional Random Fields(CRF) have been used successfully to tag sentences automatically.
- Software tools: Stanford Tagger, Ratnaparki’s MaxEnt tagger, Brill Tagger and TnT tagger .

Part-Of-Speech (POS) Tagging

Example of a tagged piece of text:

- The/DET limitations/N of/P traditional/ADJ and/CO structuralist/ADJ grammars/N should/AV be/BV clearly/ADV appreciated/V. Although/CON/COM such/ADJ grammars/N may/MV contain/V full/ADJ and/CO explicit/ADJ list/N/V of/P exceptions/N and/CO irregularities/N, they/PPN provide/V only/ADJ/CON examples/N and hints/V/N concerning/VG the/DET regular/ADJ and/CO productive/ADJ syntactic/ADJ process/N/V. Traditional/ADJ linguistic/ADJ theory/N was/BV not unaware/ADJ of/P this/DET fact/N.

Chunk Parsing

- Steve Abney pioneered the notion of chunk parsing.

  [I begin] [with an intuition]: [when I read] [a sentence], [I read it][a chunk] at a time.
- Chunks appear to correspond roughly to prosodic phrases (though nobody has really researched this form the phonology-syntax interface perspective). Strong stress will be on a chunk.
- Pauses are most likely to fall between chunks.
- Chunks are difficult to define Precisely.

Chunk Parsing

"A typical chunk consists of a content word surrounded by a constellation of function words, matching a fixed template." (Abnen 1991).

[I begin] [with an intuition]: [when I read] [a sentence], [I read it] [a chunk] [at a time]

[The girl] saw [the monkey] [with the telescope]

- By contrast, the relationship between chunks is mediated more by lexical selection than by rigid templates."(Abney 1991).

Chunk Parsing

Abney envisioned chunk parsing as cascade of (finite-state) processes:

- Word Identification (in practice mostly done via POS tagging)
- Chunk Identification (via rules)
- Attachment of Chunks (Parsing, also via rules)

Chunker

- Basically, a chunker divides a sentence into its major non-overlapping phrases and attaches a label to each.

  [NP The student] from [NP Germany] [VP was trying to phone] [NP his parents] [NP last night]

Chunker

- Chunkers differ in terms of their precise output and the way in which a chunk is defined. Many do more than just simple chunking. Others just find NPs.
- Chunking falls between tagging (which is feasible but sometimes of limited use) and full parsing (which is more useful but is difficult on unrestricted text and may result in massive ambiguity).

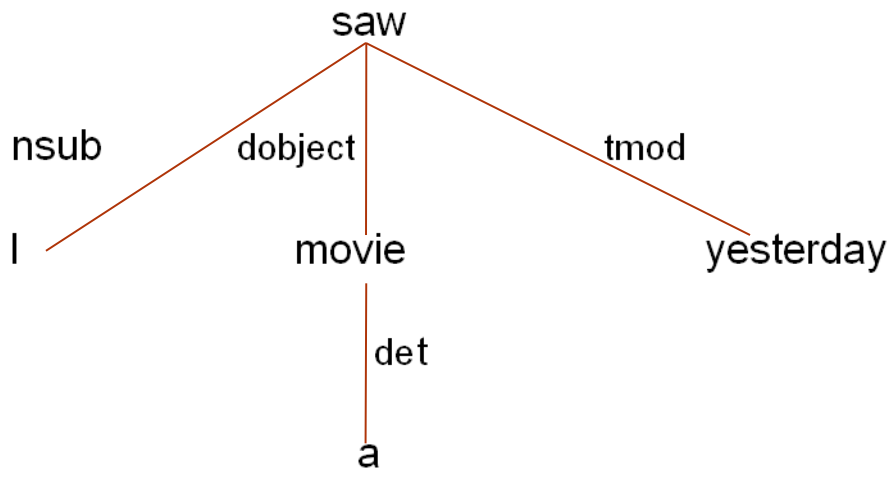
Chunker

- The structure of individual chunk is fairly easy to describe, while relationship between chunks are harder and more dependent on individual lexical properties.
- So chunking is a compromise between the currently available and the ideal processing output.
- Chunkers tokenise and tag the sentence.
- Mostly chunkers simply use the information in tags, but other look at actual words.

Dependency Parsing (DP)

- Alternative to phrase structure approach to syntactic analysis
- Identifies grammatical relations that connects one word to another
- Relation holds between governor and dependent
- More closer to semantic representation.
- Would be suitable for free word order languages such as Kannada, Tamil, Malayalam and Telugu.
- Software tools: Link parser, Stanford parser, Malt parser and Minipar.

Dependency Parsing (DP)



nsubj(saw-2, I-1)

det(movie-4, a-3)

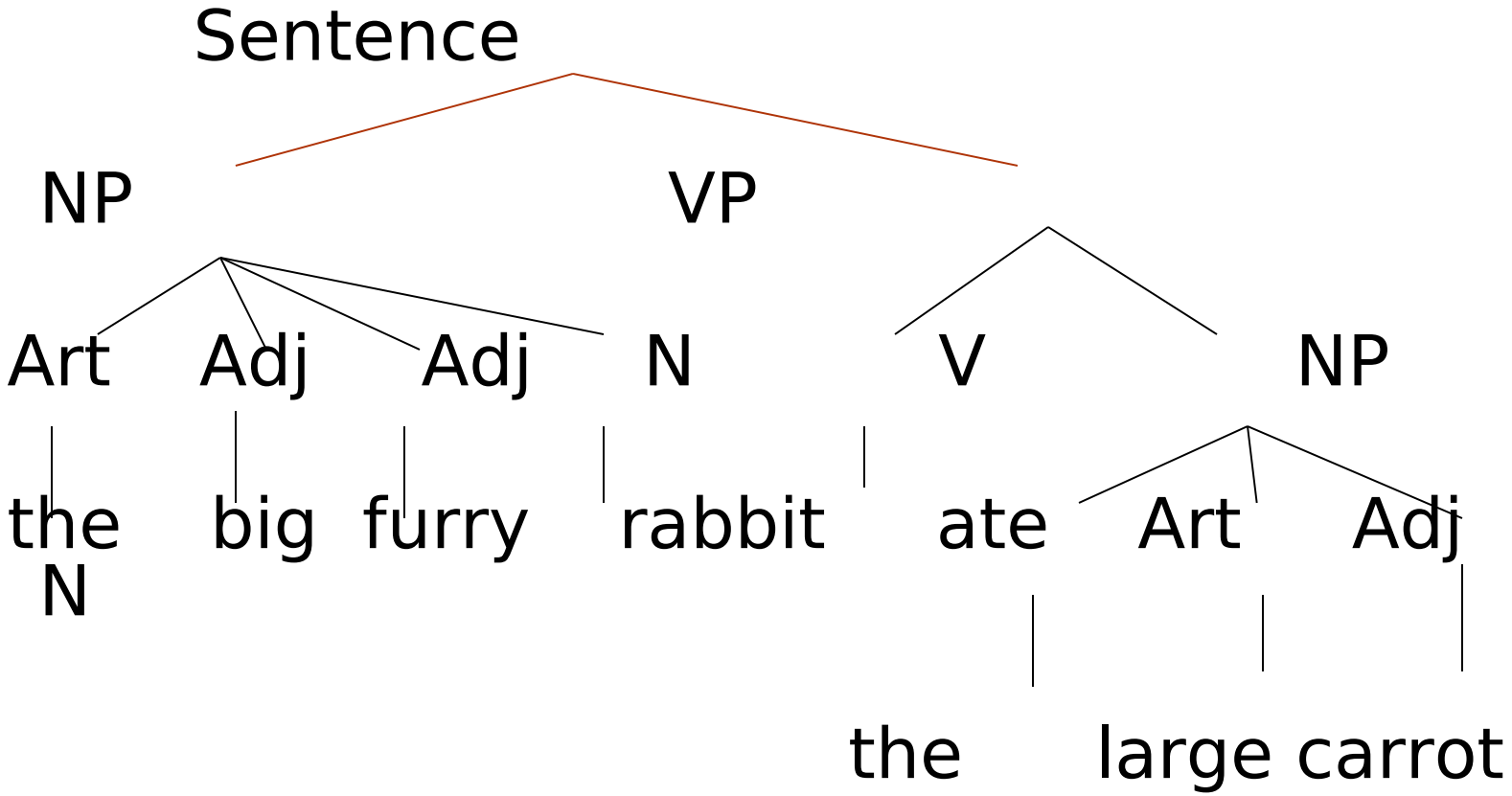
dobj(saw-2, movie-4)

tmod(saw-2, yesterday-5)

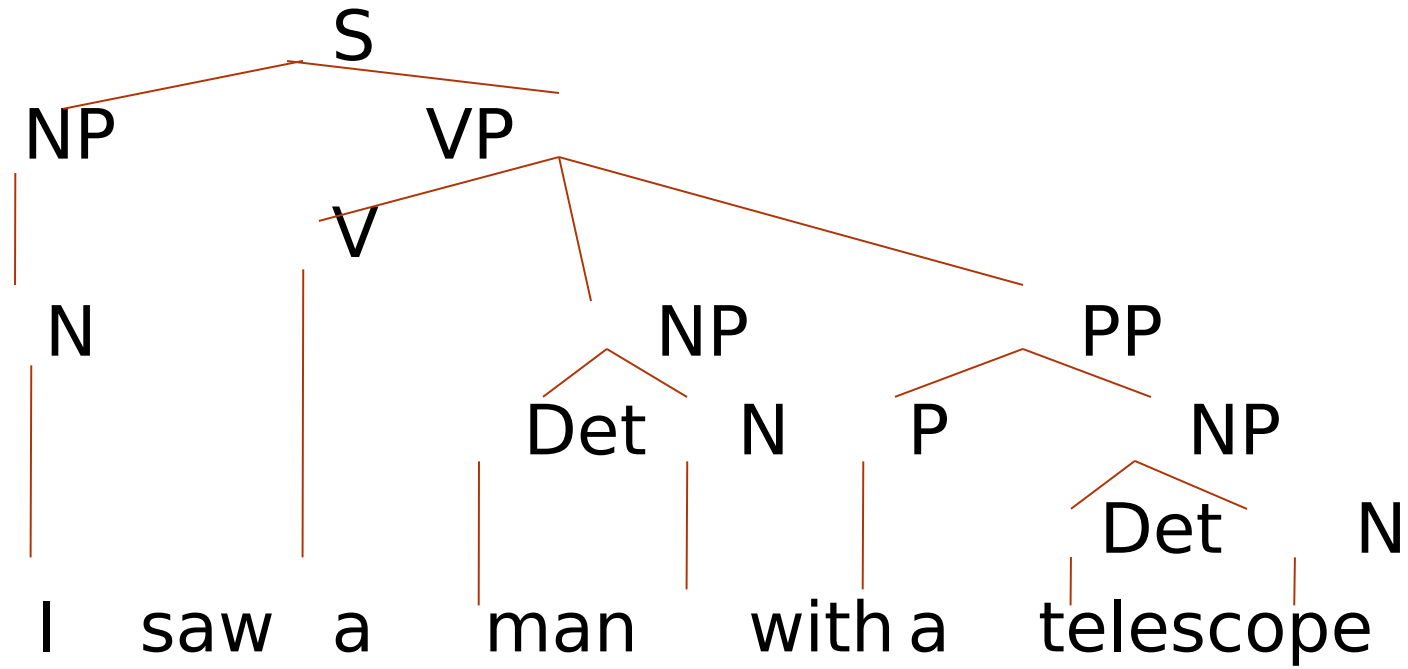
Parsing

- Parsing in natural language context is a process of recognizing input sentence and assigning syntactic structure to it.
- Attachment ambiguity: (PP-attachment) “I saw a man with a telescope”
- Parser should be able to solve ambiguities at structural level.
- Earley and CYK parsing techniques available for CFG grammars.
- Software tools: Stanford Parser, XTAG Parser from Upenn, Charniak parser, Collins parser, Malt parser etc...

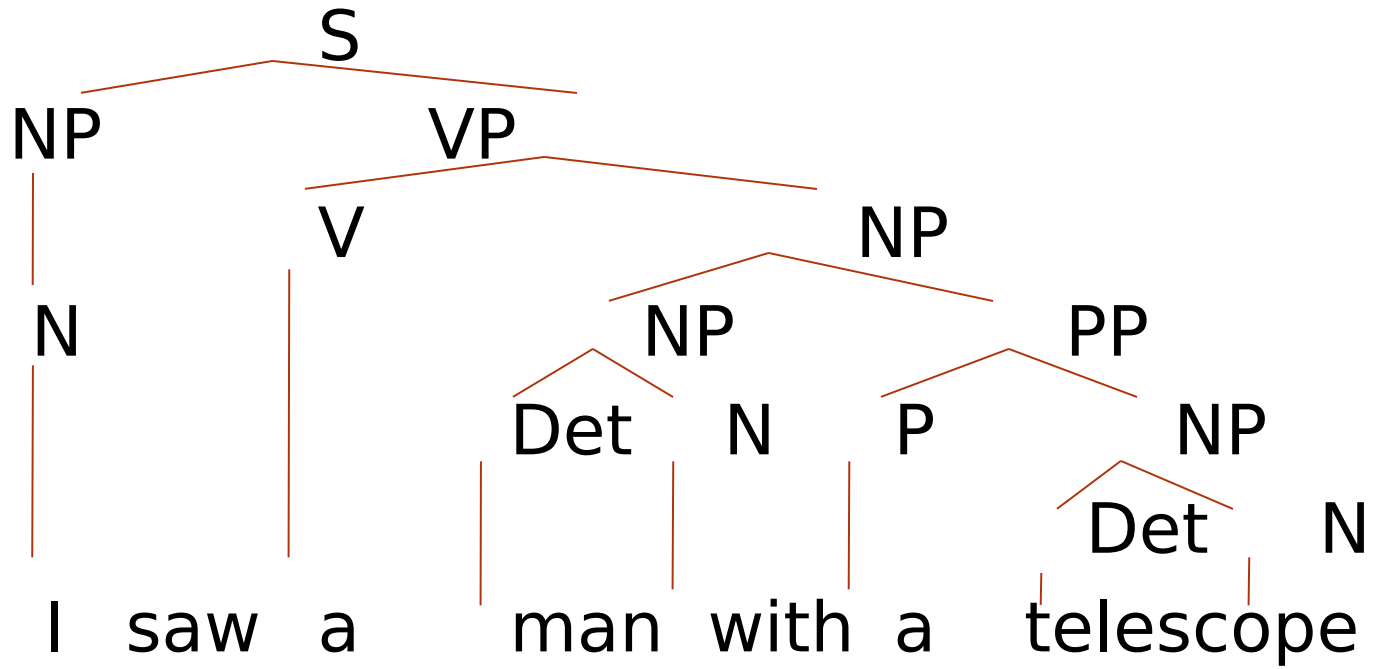
Syntactic Parsing



Parsing



Parsing



Semantic analysis

- This is the level at which most people think meaning is determined, however, as we can see in the above defining of the levels, it is all the levels that contribute to meaning.
- Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence.
- This level of processing can include the semantic disambiguation of words with multiple senses; in an analogous way to how syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished at the syntactic level.
- Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation of the sentence.

Semantic analysis

- For example, amongst other meanings, ‘file’ as a noun can mean either a folder for storing papers, or a tool to shape one’s fingernails, or a line of individuals in a queue.
- If information from the rest of the sentence were required for the disambiguation, the semantic, not the lexical level, would do the disambiguation.
- A wide range of methods can be implemented to accomplish the disambiguation, some which require information as to the frequency with which each sense occurs in a particular corpus of interest, or in general usage, some which require consideration of the local context, and others which utilize pragmatic knowledge of the domain of the document.

Semantic Processing

- Meaning can be subdivided in many ways. A simple division is between semantics, referring to the meaning of words and thus the meaning of sentences formed by those words and pragmatics, referring to the meanings intended by the originator of the words.
- Note the meaning of sentences depends not only on the meaning of the words, but also on rules about the meaning of word combinations, word orders, etc.
- For example, in the English sentence *the boy sees the girl*, we know that *the boy* does the seeing and *the girl* is seen because in English the normal word order is subject--verb--object.
- Having analysed sentence structure we try to obtain a representation of what it means.
- Different sentences with same meaning should have the same formal representation.

Semantic Processing

- For example:

  1. John gave the book to Mary.
  2. The book was given to Mary by John.

  Both could be represented formally as a structure like:

  give: action

  agent: john

  object: book

  receiver: mary
- Appropriate data structures would be used to represent this - at the simplest way just a "structure".
- It should then be possible to manipulate and extract information from this representation.
- For example, we would answer the question: ``Who gave a book to Mary?'' by looking for a structure with "mary" in the receiver field, "book" in the object field, and then returning whatever is found in the agent field.

Other Important issues

There other important issues to be addressed

1. Word Sense Disambiguation (WSD)
2. Name Entity Recognition
3. Multiword Expressions
4. Anaphora Resolution/coreferential resolution

Word sense disambiguation (WSD)

- Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a computational manner.
- WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. We introduce the reader to the motivations for solving the ambiguity of words and provide a description of the task.
- There are at least three kinds of WSD system: supervised, unsupervised, and knowledge-based approaches.

Word Sense Disambiguation

* “An empty **can**.” – “**Can** he do that?” – “**Can** it!”
* “Run **fast**” – “Stand **fast**” (antonymy).
* “Time flies like an arrow.”
* **Book** that flight
* I walked to the **bank** *(homonymy*)
* Chennai is the **capital** city of Tamilnadu *(polysemy*)
* **Cleaning fluids** can be dangerous.

Multiword Expressions

- Compounds of the type N+N, V+V, Adj+Adj, Adv+Adv, verb+auxiliary verb, N+Verbalizer, onomatopoeic compounds, etc. need to be taken together while parsing.

e.g.

Old ladies hostel

Anaphora resolution

- Anaphora resolution is the replacing of words such as pronouns, which are semantically vacant, with the appropriate entity to which they refer.

  - The city councilors refused the demonstrators a permit because they feared violence.
  - The city councilors refused the demonstrators a permit because they advocated revolution.

Anaphora resolution

Referring “back” to already mentioned entities:

* “I thought about bringing the bike, but decided to leave it at home.”
* “I thought about bringing the bike, but decided against it.”
* “I thought about bringing the bike, but the tire was flat.”

Ambiguous references:

* “I got a coffee at a place on the way here and finished it while waiting.”
* “I was going to get a coffee at a place on the way here but it was closed.”
* “I got a coffee at a place on the way here. It wasn’t as long as I thought it would be.”
* “I got a coffee at a place on the way here. It wasn’t easy.”

Named entity recognition

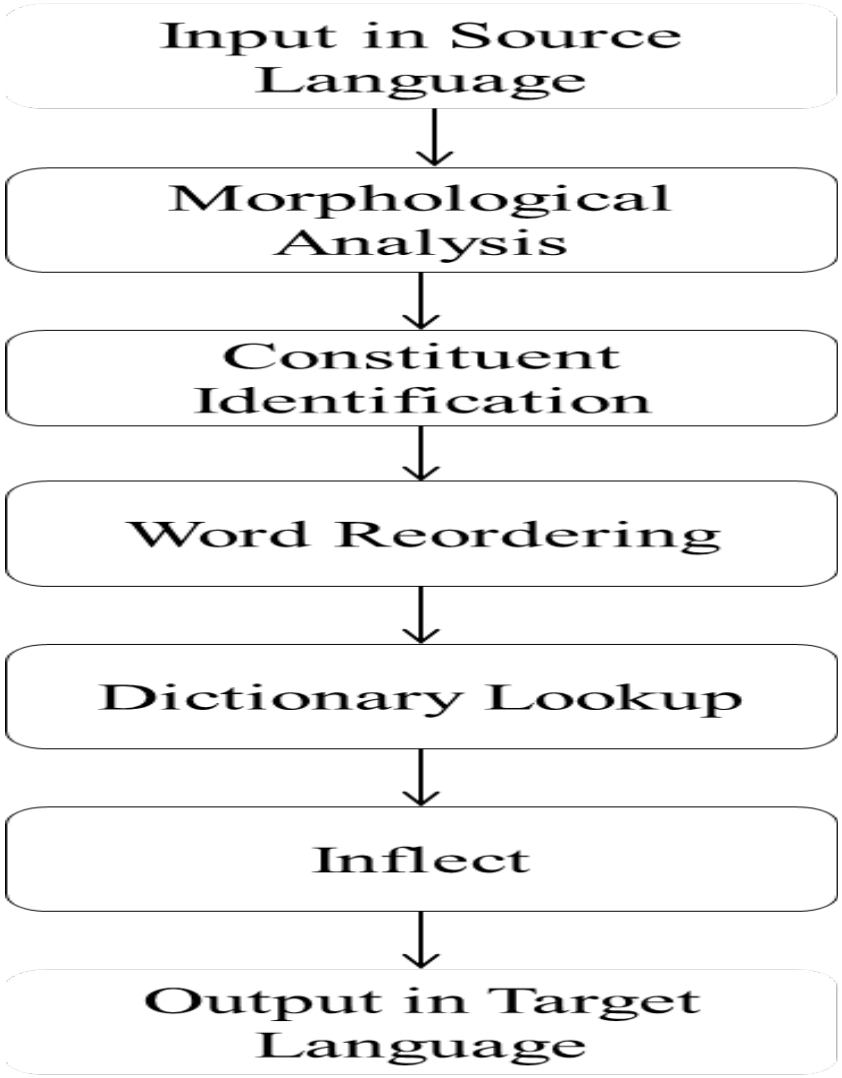
- Named entities include names of person, place, commodity, etc.
- They have to identified and separated out of other words which are found in common dictionaries.
- They need to be recognized as they undergo inflections like ordinary words.

ex.

Madurai Kamaraj University

World Heath Organization

Direct Approach



An Example to illustrate direct approach to MT

Input Sentence in English

He came late to school yesterday

After

Morphological Analysis

He come PAST late to school yesterday

Constituent Identification

<He> <come PAST> <late> <to school> <yesterday>

Word Reordering

<He> <yesterday> <to school> <late> <come PAST>

Dictionary Lookup

Vah kal ter se skuul aa_PAST

Inflect(the final translated sentence)

Vah kal ter se skuul aayaa .

वह देर से स्कूल आया

Interlingua Approach

- Interlingua approach to machine translation mainly aims at transforming the texts in the source language to a common representation applicable to many languages, using which translation of text to the target language is performed.
- Interlingua approach sees machine translation as a two stage process:
- Analysing and transforming the source language texts into a common language independent representation.
- From the common language independent form generate the text in the target language.

Interlingua Approach

- The first stage is particular to source language and doesn’t require any knowledge about the target language whereas the second stage is particular to the target language and doesn’t require any knowledge from the source language.
- The main advantage of interlingua approach is that it creates an economical multilingual environment that requires 2n translation systems to translate among n languages where in the other case, the direct approach requires n(n-1) translation systems.
- Table 2.2 has the interlingua representation of the sentence, “he will reach the hospital in ambulance”.

Interlingua Approach

An example to illustrate the Interlingua Representation

Predicate
Reach
Agent
Boy (Number: Singular)
Theme
Hospital (Number: Singular)
Instrument
Ambulance (Number: Singular)
Tense
FUTURE

Interlingua Approach

- The concepts and relations that are used known as the ontology are the most important aspect in any interlingua-based system.
- The ontology should be powerful enough that all subtleties of meaning that can be expressed using any language should be representable in the interlingua.
- Interlingua approach can be found more economical when translation is carried out with three or more languages and also the complexity of this approach gets increased, dramatically.
- This is clearly evident from the Vauquios triangle which is shown in the Figure

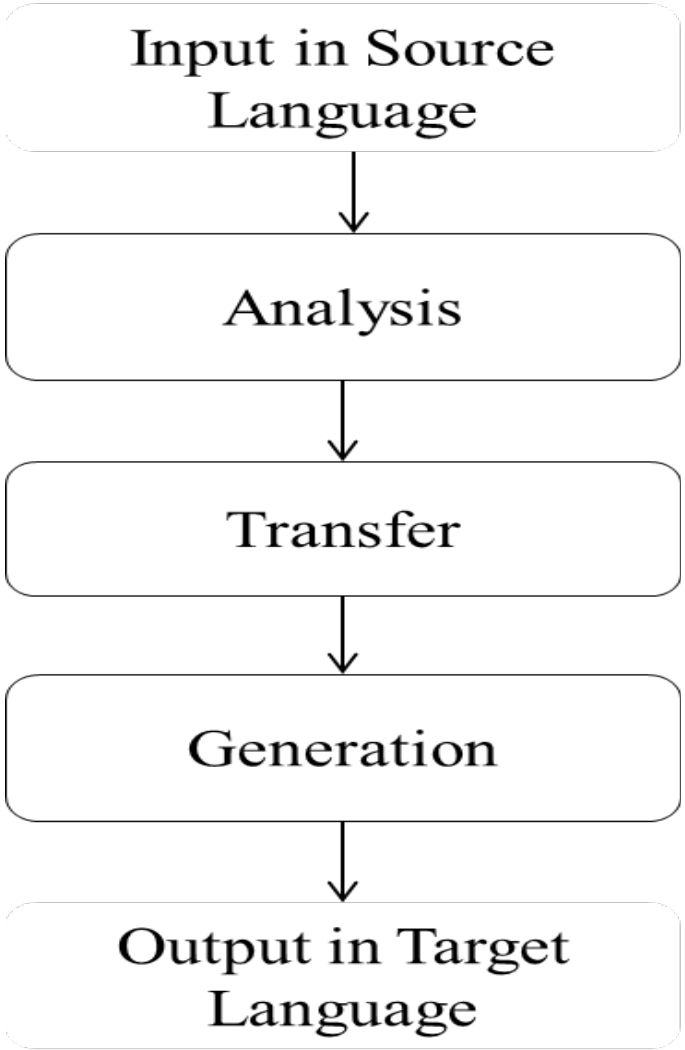
Transfer Approach

- The less determined transfer approach has three stages comprising the intellectual representations of the source and target language texts, instead of the two stages in the interlingua approach.
- The transfer approach can be done either by considering the syntactic or semantic information of the text.
- In general, transfer can either be syntactic or semantic depending on the need.

Transfer Approach

- The transfer model involves three stages: analysis, transfer, and generation.
- In the analysis stage, the source language sentence is parsed, and the sentence structure and the constituents of the sentence are identified.
- In the transfer stage, transformations are applied to the source language parse tree to convert the structure to that of the target language.
- The generation stage translates the words and expresses the tense, number, gender etc. in the target language.
- The following figure shows the block diagram of the transfer approach.

Block diagram of Transfer Approach to Machine Translation



Transfer Approach

Consider the sentence, “*he will come to school in bus*”. The table illustrates the three stages of the translation of this sentence using the transfer approach.

Table :An example to illustrate to the Transfer Approach

Input Sentence

He will come to school in bus

Analysis

<he> <will come> < to school> < in bus>

Transfer

<he> <in bus> <to school> <will come>

Generation (Output)

वह बस में स्कूल आएगा

Transfer Approach

- From the above example it will be clear that, the analysis stage of this approach produces a representation that is source language dependent and the generation stage generates the final translation from the target language dependent representation of the sentence.
- Thus, using this approach in multilingual machine translation system to translate n languages, will require ‘n’ analyser components, n(n-1) transfer components since individual transfer components are required for translation between a pair of languages for each direction and ‘n’ generation components.

Non-Linguistic Approaches

- The non-linguistic approaches are those which don’t require any linguistic knowledge explicitly to translate texts in the source language to target language.
- The only resource required by this type of approaches is data, either the dictionaries for the dictionary based approach or bilingual and monolingual corpus for the empirical or corpus based approaches.

Dictionary Based Approach

- The dictionary based approach to machine translation uses a dictionary for the language pair to translate the texts in the source language to target language.
- In this approach, word level translations will be done.
- This dictionary based approach can either be preceded by some pre-processing stages to analyse the morphological information and lemmatize the word to be retrieved from the dictionary.
- This kind of approach can be used to translate the phrases in a sentence and found to be least useful in translating a full sentence.
- This approach will be very useful in accelerating the human translation, by providing meaningful word translations and limiting the work of humans to correcting the syntax and grammar of the sentence.

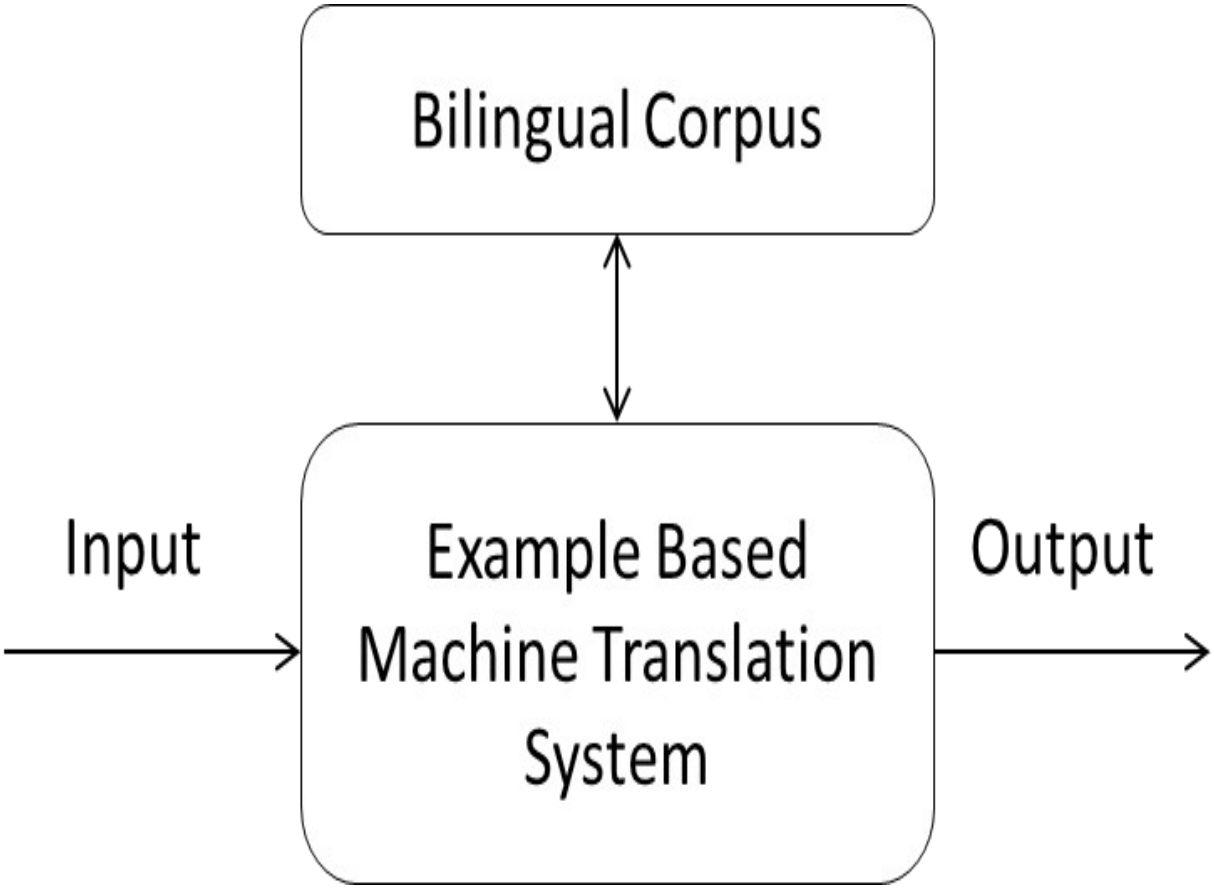
Empirical or Corpus Based Approaches

- The corpus based approaches don’t require any explicit linguistic knowledge to translate the sentence.
- But a bilingual corpus of the language pair and the monolingual corpus of the target language are required to train the system to translate a sentence.
- This approach has driven lots of interest world-wide, from late 1980s till now.

Example Based Approach

- This approach to machine translation is a technique that is mainly based how human beings interpret and solve the problems.
- That is, normally the humans split the problem into sub problems, solve each of the sub problems with the idea of how they solved this type of similar problems in the past and integrate them to solve the problem in whole.
- This approach needs a huge bilingual corpus of the language pair among which translation has to be performed. The figure given below shows the block diagram of example-based approach.
- The assumption made here is “Sentences of similar structure in the source language have a similar structure in the target language.”

Block Diagram of Example Based Machine Translation System



Example Based Approach

- In order to get a clear idea of this approach, consider the following sentence,

“He bought a home”

- Assuming that we are using a corpus that contains the following two sentence pairs:
- Table: An example to illustrate example based approach.

English	**Hindi**
He bought a pen	us ne ek kalam kharitliyaa
He has a home	uski ek ghar khai

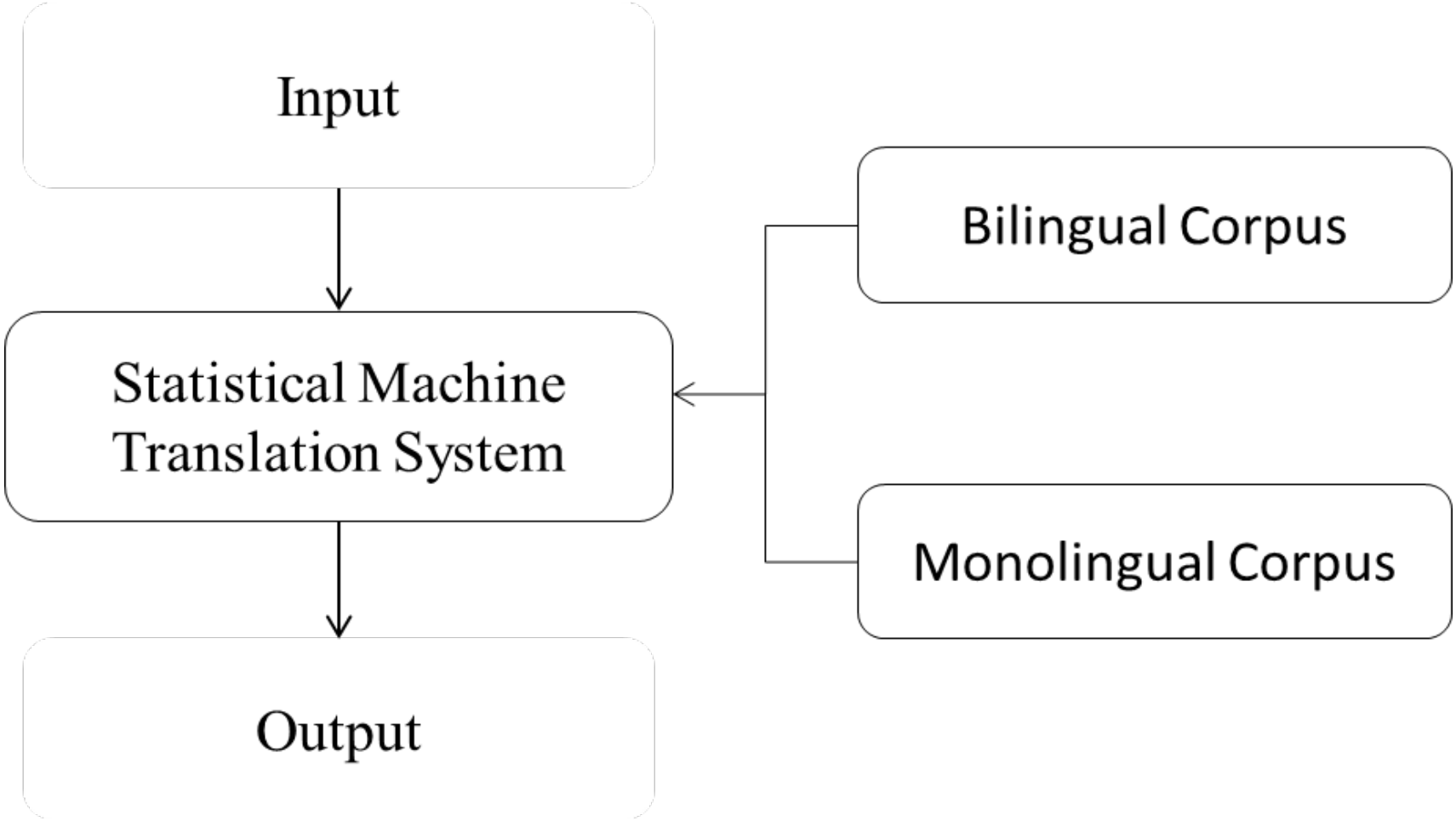
Example Based Approach

- The parts of the sentence to be translated will be matched with these two sentences in the corpus.
- Here, the part of the sentence ‘He bought’ gets matched with the words in the first sentence pair and ‘a home’ gets matched with the words in the second sentence pair.
- Therefore, the corresponding Hindi part of the matched segments of the sentences in the corpus are taken and combined appropriately.
- Sometimes, post-processing may be required in order to handle numbers, gender if exact words are not available in the corpus.

Statistical Approach

- Statistical approach to machine translation generates translations using statistical methods by deriving the parameters for those methods by analysing the bilingual corpora.
- This approach differs from the other approaches to machine translation in many aspects. Figure given below shows the simple block diagram of a statistical machine translation system.

Simple Block Diagram of Statistical Machine Translation System



Statistical Approach

- The advantages of statistical approach over other machine translation approaches are as follows:
- The enhanced usage of resources available for machine translation such as manually translated parallel and aligned texts of a language pair, books available in both languages and so on. That is large amount of machine readable natural language texts are available with which this approach can be applied.
- In general, statistical machine translation systems are language independent i.e., it is not designed specifically for a pair of language.

Statistical Approach

- Rule based machine translation systems are generally expensive as they employ manual creation of linguistic rules and also these systems cannot be generalised for other languages, whereas statistical systems can be generalised for any pair of languages, if bilingual corpora for that particular language pair is available.
- Translations produced by statistical systems are more natural compared to that of other systems, as it is trained from the real time texts available from bilingual corpora and also the fluency of the sentence will be guided by a monolingual corpus of the target language.

Statistical Approach

- This approach makes use of translation and language models generated by analysing and determining the parameters for these models from the bilingual corpora and monolingual corpus of the target language, respectively.
- Even though designing a statistical system for a particular language pair is a rapid process, the work lies on creating bilingual corpora for that particular language pair, as this was the technology behind this approach.
- In order to obtain better translations from this approach, at least more than two million words if designing the system for a particular domain and more than this for designing a general system for translating particular language pair.
- Moreover, statistical machine translation requires an extensive hardware configuration to create translation models in order to reach average performance levels.

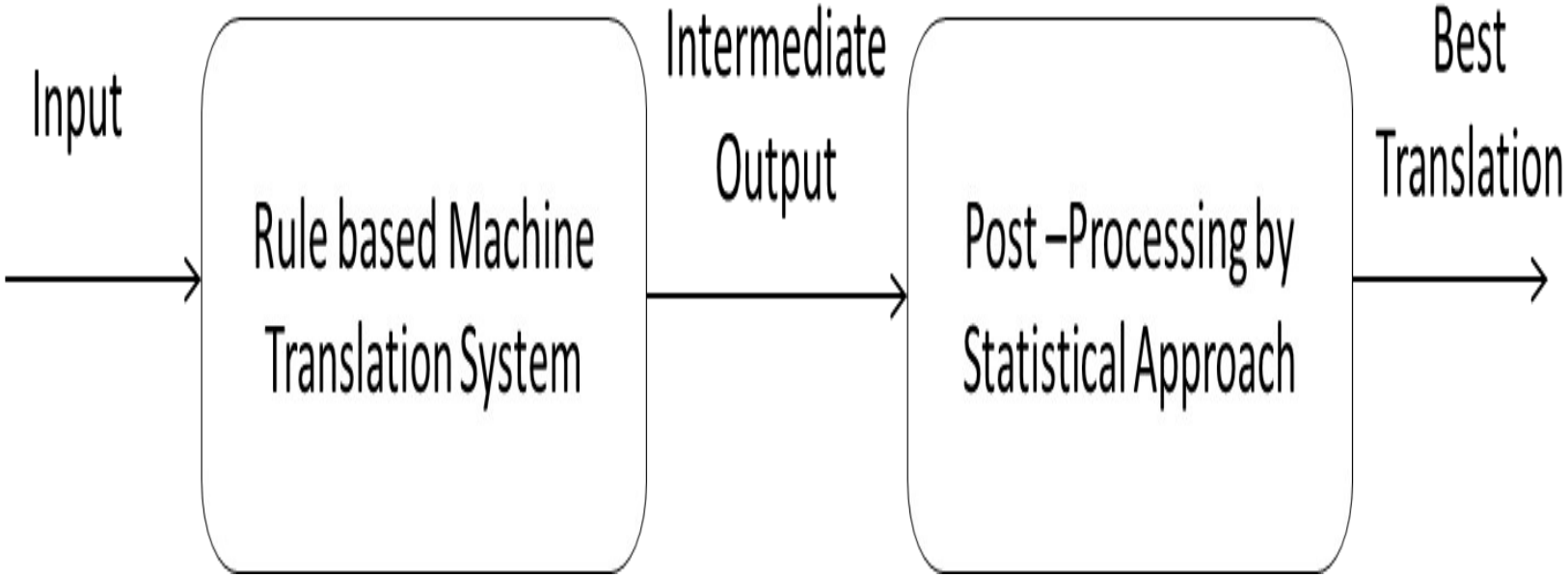
Hybrid Machine Translation Approach

- Hybrid machine translation approach makes use of the advantages of both statistical and rule-based translation methodologies.
- Commercial translation systems such as Asia Online and Systran provide systems that were implemented using this approach.
- Hybrid machine translation approaches differ in many numbers of aspects:

Rule-based system with post-processing by statistical approach

- Here the rule based machine translation system produces translations for a given text in source language to text in target language.
- The output of this rule based system will be post-processed by a statistical system to provide better translations.
- Figure given below shows the block diagram for this type of system.

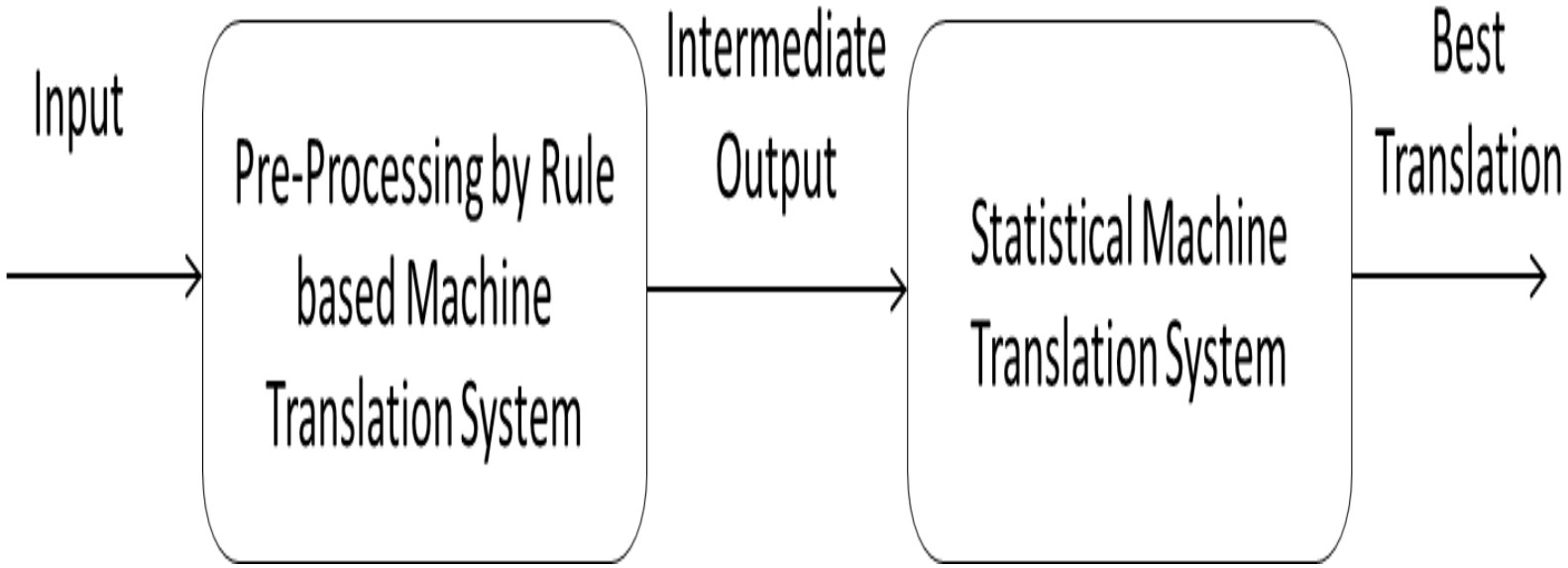
Rule-based translation system with post-processing by statistical approach



Statistical translation system with pre-processing by the rule based approach

- In this approach a statistical machine translation system is incorporated with a rule based system to pre-process the data before providing the data for training and testing.
- Also the output of the statistical system can also be post-processed using the rule based system to provide better translations. The block diagram for this this type of system is shown in Figure below.

Statistical translation system with pre-processing by the rule based approach



Machine Translation in India

- MT Research in India started in early 90's. Most of the operational and experimental systems are rule-based.
- Some important projects include:
- - Mantra by CDAC, Pune
  - Anusaaraka by IIT-Kanpur and University of Hyderabad
  - AnglaBharti by IIT Kanpur
  - Shakti by LTRC, IIIT Hyderabad
  - MaTra by CDAC, Mumbai
- IL-IL MT Consortia project: Language pairs include Malayalam-Tamil, Marathi-Hindi, Bengali- Hindi, Telugu-Hindi, Tamil-Telugu.
- E-IL MT Consortia project: Language pairs include English - Hindi, Tamil, Marathi, Oriya, Urdu, Bengali, Bodo

English-Indian languages MT

- Morphological issues: English is morphologically simple, whereas some of the Indian languages have complex morphology.
- inflectional. In some of the Indian languages especially in Dravidian group of languages (except Malayalam) the finite verbs takes gender markers.
- Syntax: English syntax follows SVO order whereas mostof the Indian languages follows SOV order.
- Semantics: Idioms and lexical ambiguities has to be resolved at this level.

General Challenges in MT

Barriers in good quality MT output can be attributed to *ambiguity* in natural languages. Ambiguity can be classified into two: **Structurally ambiguous** and **Lexically ambiguous**

Structural ambiguity:
"I saw the man on the hill with the telescope"

Lexical ambiguity:

Book that flight

I walked to the **bank** *(homonymy*)

Chennai is the **capital** city of Tamilnadu *(po/ysemy*)

Cleaning fluids can be dangerous.

English-Hindi differences

Noun cases	Hindi pospositions	English postposition or words
Nominative	0,ne	
Accusative	**ko**	
Dative	ko	to, for, at
Benefactive	ko	
Instrumental	se	with, of, by
Sociative	Ke_sat	with
Locative	Par, meM, etc.	in, on, among, to, from, with
Ablative	se	from
Genitive	kaa	of, 's

General Challenges in MT

- Barriers in good quality MT output can be attributed to ambiguity in natural languages. Ambiguity can be classified into two: Structurally ambiguous and Lexically ambiguous
- Structural ambiguity:

  "I saw the man on the hill with the telescope"
- Lexical ambiguity:

  Book that flight

I walked to the bank (homonymy)

Chennai is the capital city of Tamilnadu (polysemy)

Cleaning fluids can be dangerous.

How hard is EL-IL MT?

- The previous example illustrates the following facts
  - Dravidian languages follows SOV order as against SVO order of English
  - DLs are a free word order languages
  - DLs are agglutinative
- Difference occurs in
  - Syntactic level i.e. word ordering
  - Morphological level
- The above facts justifies the need for an efficient
  - Syntax reordering module which takes care of syntactic differences
  - Morphological generator which takes care of complex morphology of the target language.

DISCUSSION